Renewable Energy and Power Quality Journal
https://repqj.com/index.php/repqj/index
RE&PQJ, Vol. 23, No. 6, 2025

Recognition Accuracy Optimization of Power Grid Dispatching Voice Interaction System Based on Wav2Vec 2.0 and Conformer

Min Gao, Chenguang Zhu, Lei Chen, Weizhe Sun, Wengang Wang*

Pinggao Group Co., LTD., Pingdingshan, 467000, Henan Province, China
*Corresponding author's mail: wangwengang789@hotmail.com

Abstract. In view of the problem that the current power grid dispatching voice interaction system is not adaptable enough to the terminology specific to the power grid dispatching field and is easily disturbed by environmental noise, resulting in command recognition errors and missing keywords, this paper constructs a deep fusion model based on Wav2Vec 2.0 self-supervised pre-training and Conformer structure, aiming to achieve Automatic Speech Recognition (ASR) and optimize accuracy. First, based on the Wav2Vec 2.0 model, the original dispatching voice signal is self-supervised pre-trained to extract features and capture its low-level time domain and frequency domain expressions. Then, the extracted voice features are input into the Conformer structure fine-tuned by the dispatching field corpus to achieve high-precision modeling of long-distance context. Finally, the power grid professional terminology dictionary is embedded in the decoding stage, and the spectrogram enhancement and background noise synthesis mechanism are combined to achieve end-to-end joint optimization. The results showed that the accuracy, recall, and F1 score of the speech recognition model in this article were 92.3%, 89.1%, and 90.7%, respectively, with an average of Word Error Rate (WER), Character Error Rate (CER), Weighted WER were 10.8%, 5.7%, and 13.8%, respectively; The F1 score for term recognition reached 90.7%; The recognition rate of Top-3 is above 0.75, and the complete recognition rate of instructions reaches 84.6%. Under extreme low signal-to-noise ratio conditions of -5dB, its WER control is 42.1%. The conclusion shows that the method proposed in this paper can effectively improve the accuracy and scene adaptability of ASR, provide reliable support for high-precision voice interaction in power grid scheduling, help improve the safety and reliability of power facility operations, and reduce work delays caused by misoperation or poor communication.

Key words. Power Grid Dispatching, Voice Recognition Optimization, Wav2Vec 2.0 Model, Conformer Model, Feature Learning

1. Introduction

With the rapid development of power systems and intelligent dispatching technology, the dispatching and operation mode of power grids has changed from manual operation to intelligent human-computer interaction and voice control [1,2]. As an important interface for efficient human-computer collaboration, voice interaction systems play an increasingly important role in dispatching automation, job instruction execution, and emergency processing [3]. In recent years, ASR technology, especially hybrid acoustic model frameworks, have made significant progress driven by deep learning. But its complexity and computational cost limit its application scope. The currently widely used general ASR model still faces the problems of low recognition precision, easy loss of keywords, and semantic errors in power dispatching scenarios facing special vocabulary, fixed instruction format, and multi-source complex voice input, which seriously affects the stability and reliable operation of the system in the actual dispatching environment. Improving the ASR system's modeling ability for dispatching instruction language and its adaptability in practical applications are the core issues that need to be urgently solved to realize the in-depth application of voice interaction technology in the power industry.

Wav2Vec2.0 can effectively extract contextual information in voice without manual labeling and has good transfer learning ability [4]. Conformer combines the local modeling capability of the Convolutional neural network (CNN) with the global attention mechanism of the Transformer. It has great advantages in temporal modeling and semantic representation and is particularly suitable for processing long-distance related complex voice instructions [5,6]. This paper combines Wav2Vec 2.0 with the Conformer to study the optimization method of power grid dispatch ASR. Through self-supervised learning, the unlabeled voice features are deeply encoded, and the Conformer network is integrated to improve the semantic model of remote instructions. At the same time, the language model and joint optimization mechanism built based on the dispatch corpus are applied to improve the system's recognition precision of industry terms and instructions, which has important theoretical value and practical significance for promoting the evolution of human-machine collaboration in smart grids and

improving the efficiency and response capabilities of power grid operations.

This article aims to optimize the recognition accuracy of the power grid dispatch voice interaction system based on Wav2Vec 2.0 and Conformer model. By improving the existing Wav2Vec 2.0 and Conformer architecture, the recognition accuracy of power grid dispatch related terms and instructions can be improved, ensuring that the system can maintain high-precision recognition performance under different environmental noise conditions. The innovation of this article lies in: 1) fully utilizing the self supervised learning ability of Wav2Vec 2.0, combined with the Conformer temporal modeling mechanism, to achieve joint optimization of speech feature extraction and contextual information acquisition. Based on Wav2Vec 2.0, rich feature representations are extracted from a large amount of unlabeled audio data, and combined with the local and global information processing capabilities of the Conformer model, attention to local details is retained while capturing long-distance dependencies, achieving accurate understanding of the context of the overall dialogue, as well as accurate recognition of individual commands and keywords, improving its recognition accuracy and robustness in power grid scheduling; 2) the professional-oriented attention mechanism is applied to enable it to have the ability to dynamically focus on keywords, improving its recognition accuracy and recall rate; 3) in view of the complex and changeable characteristics of power grid dispatching scenarios, the noise adaptation mechanism based on pre-training characteristics and structural optimization is used to achieve recognition capabilities under extremely low signal-to-noise ratio conditions, thereby improving the system's stability and reliability.

2. Related Works

The existing research on power grid dispatching ASR systems mainly focuses on command automation parsing and voice-to-text accuracy improvement [7,8]. Zhihua Wang proposed a power dispatching voice interaction model based on deep convolutional Generative Adversarial Networks (GANs). Based on GAN, convolutional layers and conditional information were added, and conditional information was used to generate high-quality voice. The results showed that compared with methods such as spectrum subtraction, the proposed model could achieve better voice enhancement [9]. Chen Lei proposed an ASR model based on BERT (Bidirectional Encoder Representations from Transformers). According to the characteristics of power grid dispatching language, semantic features, keyword features, and named entity features were extracted from dispatching sentences to improve the adaptability of the model to power grid dispatching language. The example results showed that the proposed model had a relatively obvious advantage in ASR accuracy [10]. To make the professional vocabulary in the field of power grid dispatching computable, Hao Feng proposed a word to vector technology based on historical corpus in the dispatching field. The actual example analysis results showed that this method improved the accuracy of ASR technology in the field of power grid dispatching [11]. To verify the correctness of dispatching instructions, Sun Lili proposed to build a dispatching authentication system based on voiceprint recognition, using CNN to extract voiceprint identity features from short-term voice and integrating the identity authentication method based on dynamic passwords to solve the problem of incorrect dispatching instructions in dispatching scenarios. The results showed that the proposed method system could efficiently respond to user requests and improve the safety and quality of dispatching work [12]. Existing research has achieved the recognition and response of basic commands in specific scenarios, but most systems rely on general ASR models, lack end-to-end modeling, and still cannot meet the high reliability requirements in terms of long-distance dependency understanding.

The mature development of Wav2Vec 2.0 and Conformer provides more possibilities for end-to-end modeling and long-distance dependency understanding [13-15]. Wav2Vec 2.0 and Conformer models are powerful tools for handling speech recognition tasks. The professional terminology and commands in the power industry are highly specialized and complex. Wav2Vec 2.0 can extract rich acoustic features from unlabeled data through a self supervised learning mechanism [16,17]. The Conformer model integrates CNN and self attention mechanisms, making it very suitable for handling complex dialogue scenarios in power grid scheduling [18,19]. Compared with other technologies, choosing Wav2Vec 2.0 and Conformer not only maintains high performance but also has relatively low computing resource requirements. Using them as infrastructure not only saves development time but also reduces project risks. Zhao Jing developed and analyzed a series of wav2vec pre-trained models for ASR of low-resource languages, used phoneme-level recognition tasks in fine-tuning, and extracted similarities from different transformer layers. The pre-trained representations were applied to end-to-end and hybrid systems, and the good performance of the proposed model was verified through experimental analysis [20]. Deng Bin proposed an ASR technology based on the Conformer model, applied a convolutional module into the Transformer model to improve the model's ability to learn subtle features, and then input the power grid dispatch language for feature extraction. Finally, through experimental verification, the proposed model had high accuracy in power grid dispatch ASR, and the word error rate on the validation set was reduced by 11.23% and 21.76%, respectively [21]. Sang Jiangkun proposed a compression optimization strategy for an end-to-end ASR model based on Conformer, which adopted three compression optimization strategies: model quantization, structured pruning based on weight channels, and singular value decomposition. At the same time, the model quantization was improved, and these strategies were combined to test on different devices. Compared with the baseline, its long-distance dependency understanding error was less than 3% [22]. Existing research has effectively enhanced the ability to model end-to-end and capture the semantics of long sequences. However, in dealing with variable voice speed, terminology ambiguity, and noise

interference in dispatching scenarios, existing models still face the problem of insufficient adaptability to exclusive terminology and voice style.

This article draws on the successful applications of Wav2Vec 2.0 and Conformer models, and conducts targeted optimization based on them. The Wav2Vec 2.0 model proposed by Facebook AI Research is adopted as the basic architecture, and feature representations are extracted from a large amount of unlabeled audio data through self supervised learning to improve the performance of speech recognition tasks; Considering the complex acoustic conditions in the power dispatch environment, the Conformer model was integrated into the system design, combining the advantages of CNN and self attention mechanism to more effectively capture long-distance dependencies and local detail information.

3. Optimization Design of Power Grid Dispatching Voice Recognition Based on Wav2Vec 2.0 and Conformer

A. Architecture of Power Grid Dispatching Voice Interaction System

The architecture of the power grid dispatch voice interaction system is shown in Figure 1:

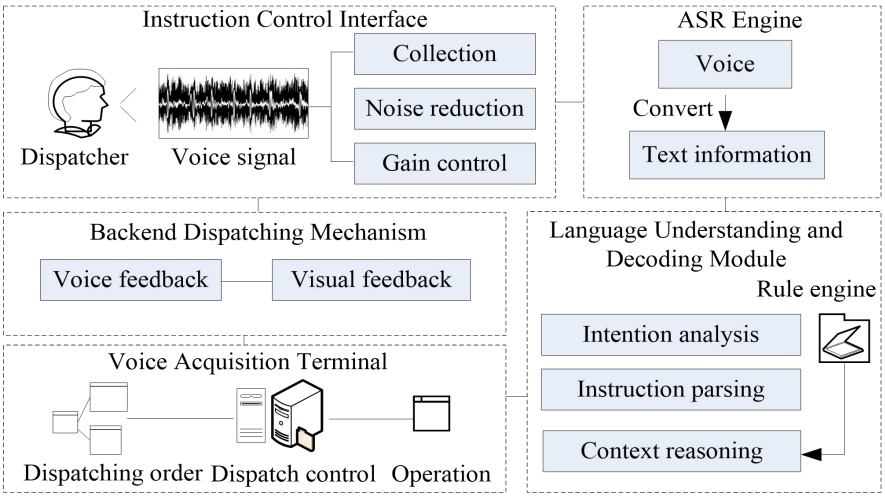

Figure 1. Architecture of power grid dispatching voice interaction system.

From the overall architecture of Figure 1, the power grid dispatch voice interaction system consists of five modules: voice acquisition terminal, ASR engine, language understanding and decoding module, instruction control interface module, and backend scheduling mechanism. The voice acquisition terminal mainly collects the voice signals sent by dispatchers and improves the voice quality through preprocessing measures such as noise reduction and gain control; The ASR engine is based on this and uses a joint model based on wav2vec 2.0 and Conformer as the ASR engine; In the language understanding and decoding module, the BERT model is used as the core component of semantic understanding. The semantic understanding part is mainly responsible for semantic parsing of text content and combining it with a rule engine to achieve semantic inference of text content; The instruction control interface is responsible for transmitting the parsed instructions to the backend scheduling information system and initiating the relevant workflow; The backend scheduling mechanism provides feedback through instructions. The speech recognition engine and language understanding module work together through a unified interface. After the speech recognition engine converts the input speech into text, the language understanding module performs semantic parsing on it and generates structured scheduling instructions. Under normal workflow, dispatchers wear or carry audio devices, and the system converts voice information into structured text, which corresponds to the semantic instructions of the power grid operation control system, thus achieving closed-loop management.

B. Wav2Vec 2.0 and Conformer Fusion Model

The current system has limited ability to capture and learn complex voice features, which easily causes key instruction content to be intercepted or misjudged. The coverage of power grid-specific vocabulary is insufficient, resulting in low recognition precision of professional terms, which seriously restricts the accurate execution of dispatch instructions. This paper aims to improve the accuracy and robustness of voice interaction, integrates Wav2Vec2.0 and Conformer structure, and establishes a recognition optimization framework that integrates feature encoding-structure modeling-professional semantic adaptation-robustness enhancement. Based on the logical integrity of the entire system, the analysis is carried out from three main modules: voice signal input and preprocessing, self-supervised feature encoding and structure modeling, and domain knowledge adaptation and robustness enhancement.

1) Voice Signal Input and Preprocessing

This paper takes the voice dialogue of dispatchers in the interactive system and the power grid dispatch voice data

collected by the sampling device of the dispatch terminal as input. Since they are very different in sampling rate, number of channels, and file format, in order to make the input data better adapt to the standardization requirements of the pre-training model, the original voice format is uniformly converted and unified into a mono and WAV format.

Since the training performance of Wav2Vec2.0 is best at a sampling rate of 16kHz, all audio signals are resampled to 16 kHz and frequency converted using FIR (Finite Impact Response) anti-aliasing filters. Its parameter design is shown in Table 1:

Table 1. Filter parameter design.

Parameter	Specification	Description
Filter type	Kaiser window FIR filter	Anti-aliasing filtering to avoid spectral distortion
Original sampling rate range	8kHz-48kHz	Source device sampling rate distribution
Target sampling rate	16kHz	Consistent with the training environment of the main model
Passband cutoff frequency	6.8kHz	Ensuring that the main frequency band of the voice is fully preserved
Stop-band attenuation	≥ 65dB	Ensuring effective suppression of high-frequency noise

In view of the large amount of non-task information such as silent waiting and background conversation in the dispatch call, the silent segments are detected and cropped. The short-time mean energy and zero-crossing rate joint gating algorithm are used to realize the silent segment detection. Given a voice frame $x_i(n)$, the definition of its short-term energy E_i is shown in formula (1):

$$E_i = \sum_{n-1}^{N} x_i(n)^2 \quad (1)$$

Its zero crossing rate Z_i is shown in formula (2):

$$Z_i = \frac{1}{2N}\sum_{n=1}^{N-1}\left|\mathrm{sgn}\left(x_i(n)\right) - \mathrm{sgn}\left(x_i(n+1)\right)\right| \quad (2)$$

The silence thresholds are set to T_E and T_Z, marked as a silent segment when the frame length is more than 300 milliseconds when $E_i < T_E$ and $Z_i < T_Z$, and cropped. Table 2 lists the silence detection parameter settings:

Table 2. Silence detection parameter settings.

Parameter	Specification	Description
Frame length	25ms	Standard frame length
Frame shift	10ms	Prevent information loss
T_E	20% of the global energy mean	Dynamically adapt to different voice intensities
T_Z	0.08	Standard for silent and weak voice boundary recognition
Cut minimum length	300ms	Avoid accidentally cutting non explicit voice content

Based on silence detection, spectral subtraction and Wiener filter are used to achieve two-stage denoising. In the first stage of spectral subtraction, in the frequency domain, the voice signal of each frame is expressed as formula (3) [23,24]:

$$Y(k) = S(k) + N(k) \quad (3)$$

In formula (3), $Y(k)$ represents the observed spectrum; $N(k)$ represents the estimated noise spectrum; $S(k)$ represents the pure voice spectrum. By estimating $N(k)$ and subtracting it from $Y(k)$, the initial denoised spectrum $\hat{S}(k)$ is obtained.

Use the Minimum Statistics method to estimate background noise. Use the minimum value of the short-time spectrum to approximate the power spectral density of background noise. Assuming the frequency spectrum of the input signal is $X(k)$, the background noise $N(k)$ can be estimated by formula (4):

$$N(k) = \min_{t\in[t_0,t_o+T]}\left|X(k,t)\right|^2 \quad (4)$$

The speech signal $S(k)$ is obtained by subtracting background noise from the total signal, as shown in formula (5):

$$S(k) = X(k) - \hat{N}(k) \quad (5)$$

The second stage introduces the Wiener filter gain function $G(k)$, as shown in formula (6):

$$G(k) = \frac{\left|S(k)\right|^2}{\left|S(k)\right|^2 + \left|N(k)\right|^2} \quad (6)$$

By smoothing the spectrum, the impact of unstructured background noise on the voice signal is reduced without

affecting the overall integrity of the voice.

After the denoising is completed, all segments are screened for effectiveness. Segments below 0.5 seconds, fuzzy pronunciation areas, and abnormal spectral energy segments are removed, and only valid sentences with real dispatching semantics are retained. On this basis, the dynamic range normalization of Root Mean Square (RMS) is used to solve the problem of training gradient fluctuations caused by different volumes, as shown in formula (7):

$$x_{\text{norm}}(n) = \frac{x(n)}{\sqrt{\frac{1}{N}\sum_{n=1}^{N} x(n)^2 + \epsilon}} \quad (7)$$

In formula (7), $x_{\text{norm}}(n)$ is the normalized signal value, ϵ is a decimal to avoid division by zero, and the value is 10^{-9}.

2) *Self-supervised Feature Encoding and Structural Modeling*

After completing the preprocessing of the voice signal, the standardized voice signal is input into the feature modeling module. In view of the lack of unlabeled data, the supervised learning acoustic model is difficult to generalize. The professional terminology of power grid dispatching voice is dense and complex. The Wav2Vec 2.0 model is used for self-supervised feature learning to achieve high-level semantic expression of the original voice from both the time domain and the frequency domain, reducing the dependence on a large number of annotations. Combined with the contextual semantic modeling of the Conformer structure, taking into account both local acoustic characteristics and global context information, a two-stage semantic alignment mechanism is adopted to achieve accurate recognition of power grid dispatching sentences with system-level semantics and long-range dependencies, as shown in Figure 2.

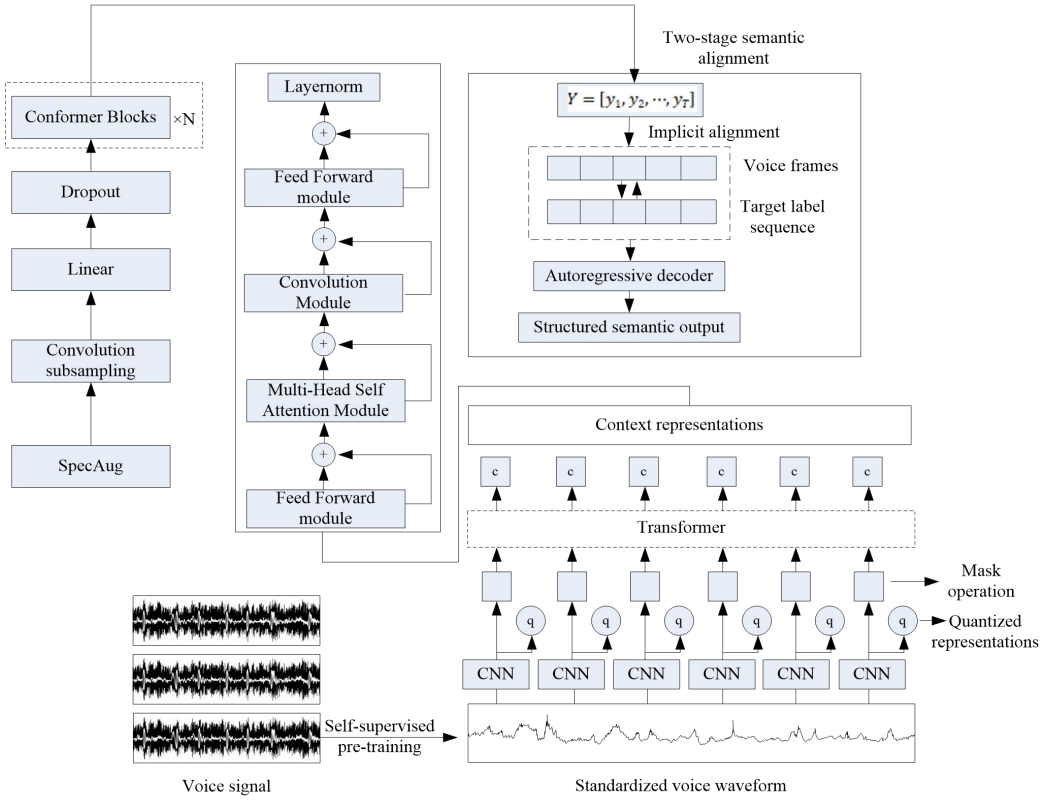

Figure 2. Self-supervised feature encoding and structural modeling mechanism.

In Figure 2, the self-supervised voice representation learning based on Wav2Vec2.0 learns the high-level semantic features of the input voice signal. Wav2vec 2.0 is responsible for extracting high-dimensional feature representations from raw speech signals, which are then input into Conformer for further time series modeling. Feature fusion modules are introduced in each layer of Conformer to concatenate the features extracted by wav2vec 2.0 with the intermediate features of Conformer. Wav2Vec2.0 consists of a feature encoder, a context network, and a vector quantization module [25]. This paper adopts the Wav2Vec 2.0 BASE model that has been pre-trained on large-scale general voice data and fine-tunes it on the local power grid dispatch voice corpus to achieve task adaptation. In the model selection, this article chose the BASE version of Wav2Vec 2.0, which contains about 95M parameters. Compared with LARGE and XLS-R, its computational complexity is significantly reduced, making it more suitable for deployment in power grid scheduling scenarios.

First, the feature encoder receives the normalized voice waveform $x \in \mathbb{R}^{\mathrm{T}}$ (T represents the number of time steps) and inputs it into a one-dimensional one-dimensional convolutional network. Then, the signal

is converted into a potential frame-level feature expression, as shown in formula (8) [26]:

$$z = f_{\text{enc}}(x) \in \mathbb{R}^{T' \times d} \quad (8)$$

In formula (8), f_{enc} is the feature encoder, $T' \ll T$, and d is a feature dimension. This processing can capture the local spectral pattern of voice and provide input for subsequent context modeling.

On this basis, the vector quantization module is used to discretize some z values $q \in \mathbb{R}^{T' \times d}$, and the Gumbel-Softmax mechanism is used to approximate the z values to construct a pseudo-supervised target [27,28].

During the learning process, a mask operation is used to replace part of z_t with a masked tag z_t^{mask}, which is then input into the Transformer context network to output a context vector expression, as shown in formula (9) [29-31]:

$$c_t = f_{\text{ctx}}\left(z_t^{\text{mask}}\right) \in \mathbb{R}^d \quad (9)$$

In formula (9), f_{ctx} is the context network. To learn effective context dependencies, based on the contrastive learning objective function, c_t and the true quantized expression q_t have the greatest similarity, while the negative sample at other time steps has the smallest similarity. Its loss function is defined as formula (10):

$$\mathcal{L}_{\text{contrastive}} = -\sum_{t \in M} \log \frac{\exp\left(\text{sim}\left(c_t, q_t\right)/\kappa\right)}{\sum_{n \in N_t} \exp\left(\text{sim}\left(c_t, q_t\right)/\kappa\right)} \quad (10)$$

In formula (10), $\text{sim}(\cdot)$ is the cosine similarity; κ represents the temperature parameter; M represents the masked position set; N_t represents the negative sample set. Self-supervised learning is then used to extract the dynamics of voice sequences, pronunciation patterns, and contextual information of pronunciation units in the unlabeled case, thereby enhancing the robustness to non-standard pronunciation, intonation changes, and stress changes.

Based on the Wav2Vec 2.0 pre-training, the high-dimensional context vector expression c_t output by it is further modeled to extract the structured instructions and semantic hierarchy of the dispatching voice. Based on Conformer, by integrating the local modeling of convolution and the global modeling of Transformer, the logical relationship between hidden instructions and the semantic features of the task are extracted from the voice data.

Conformer uses Conformer Block as the basic building block. It adds a lightweight convolution module to the standard Transformer framework and decomposes the Feedforward Neural Network (FNN). Its specific structure is:

(1) Feedforward network module

Using the residual structure, the FNN is divided into two parts, the front and the back, and each part contributes half to the residual. The input is set to x_{in}, as shown in formula (11):

$$x_1 = x_{\text{in}} + \frac{1}{2} \cdot FNN_1\left(x_{\text{in}}\right) \quad (11)$$

In formula (11), x_{in} is the input feature vector. FNN_1 is the feedforward submodule of the front part.

(2) Multi-Head Self-Attention (MHSA)

Relative positioning encoding is applied to enhance the modeling ability of the context within the voice framework,as shown in formula (12):

$$x_2 = x_1 + MHSA\left(\text{LayerNorm}\left(x_1\right)\right) \quad (12)$$

In formula (12), x_1 is the weighted sum of the residuals of the first feedforward module and the input; x_2 is the weighted sum of the second feedforward module and the input residual.

(3) Convolutional module

Gated linear units and depthwise separable convolutions are used to capture local correlations and improve the model's ability to model short-term structures.

For the gating mechanism, as shown in formula (13):

$$z_t = GLU\left(\text{Conv1D}\left(x_2\right)\right) \quad (13)$$

Batch standardization and activation are shown in formula (14):

$$\xi = BN\left(\text{DepthwiseConv}\left(z_t\right)\right) \quad (14)$$

Swish activation and output projection are shown in formula (15):

$$x_3 = x_2 + \text{ConvOut}\left(\text{Swish}\left(z_t\right)\right) \quad (15)$$

In formulas (13)-(15), Conv1D is a standard one-dimensional convolution used for local context modeling; GLU is a gated linear unit; z_t is the intermediate feature output by the convolution module; DepthwiseConv is a depthwise separable convolution that slides only within each channel; BN is batch standardization; x_3 is the sum of the output and input residuals of the convolution module; ConvOut is the

linear projection used to restore the output of the convolution module to the input dimension.

(4) Second feedforward network and normalized output, as shown in formula (16) and formula (17)

$$x_4 = x_3 + \frac{1}{2} \cdot FNN_2\left(x_3\right) \quad (16)$$

$$y = \text{LayerNorm}\left(x_4\right) \quad (17)$$

In formula (16), FNN_2 is the feedforward submodule in the latter part; x_4 is the final feature. On this basis, multiple Conformer Blocks are stacked together to form a deep acoustic representation network. The output sequence $Y = \left[y_1, y_2, \cdots, y_T\right]$ expresses the complex temporal structure and its semantic dependence on the input voice, and on this basis, the input voice is analyzed for association.

Based on the two-stage semantic alignment strategy, structured semantic units are aligned and recognized. First, the CTC (Connectionist Temporal Classification) loss function $\mathcal{L}_{\text{CTC}}$ is used to construct an implicit alignment between the input and target label sequences to solve the problem of inconsistency between voice and text, as shown in formula (18):

$$\mathcal{L}_{\text{CTC}} = -\log p\left(y \middle| X\right) = -\log \sum_{\pi \in \mathcal{B}^{-1}(y)} p\left(\pi \middle| X\right) \quad (18)$$

In formula (18), $\mathcal{B}$ is the mapping function between compressed repeated labels and blank characters, and π represents the possible path sequence. $p\left(y \middle| X\right)$ is the probability of generating the target label sequence y for a given speech input X. $p\left(\pi \middle| X\right)$ is the probability of the path sequence π.

Aligned at the frame level, the autoregressive decoder with attention mechanism is applied to generate structured semantic output step by step. According to the output $Y = \left[y_1, y_2, \cdots, y_T\right]$ of the encoder, the decoder predicts the structured text sequence $\hat{S} = \left[s_1, s_2, \cdots, s_N\right]$. Its attention weight calculation $\alpha_{t,i}$ is expressed as formula (19) [32]:

$$\alpha_{t,i} = \frac{\exp\left(q_t^{\top}, k_i\right)}{\sum_j \exp\left(q_t^{\top}, k_j\right)} \quad (19)$$

In formula (19), k_i is the key vector corresponding to the i-th time step, and q_t is the decoder query vector for the current time step; The context vector update is expressed as formula (20):

$$c_t = \sum_i \alpha_{t,i} \cdot y_i \quad (20)$$

In formula (20), c_t is the weighted sum obtained from the encoder at the t-th time step.The output word prediction is expressed as formula (21):

$$P\left(s_t \middle| s_{<t}, Y\right) = \text{soft}\max\left(W_o\left[q_t; c_t\right]\right) \quad (21)$$

In formula (21), $P\left(s_t \middle| s_{<t}, Y\right)$ is the probability of predicting the next word s_t based on the previous output $s_{<t}$ and encoder context Y. W_o is the weight matrix of the decoder output layer.

3) Domain Knowledge Adaptation and Robustness Enhancement

For multiple domain-specific instructions, device names, and symbolic words such as “main transformer”, “decoupling”, “busbar”, “220 kV”, and “double-circuit switching” in the power grid dispatching instructions, the professional terminology dictionary is embedded in the decoding process. By correcting the probability space of the decoder language model, the recognition probability and priority of key terms are improved.

The set of professional terms is defined as $\mathcal{D} = \left\{\omega_1, \omega_2, \cdots, \omega_k\right\}$, and the conditional probability distribution of the decoder output vocabulary y_t is expressed as formula (22):

$$P\left(y_t \middle| y_{<t}, X\right) = \text{soft}\max\left(W_o\left[q_t; c_t\right]\right) \quad (22)$$

In formula (22), y_t is the predicted word for the t-th time step.

During the decoding process, before the softmax calculation, for all $y_t \in \mathcal{D}$, the term weighted bias term is applied, as shown in formula (23) [33,34]:

$$P'\left(y_t \middle| y_{<t}, X\right) = \text{soft}\max\left(W_o\left[q_t; c_t\right] + \delta_t\right) \quad (23)$$

In formula (23), $P'\left(y_t \middle| y_{<t}, X\right)$ is the final prediction probability after introducing the term bias term. δ_t is the weighted bias term for professional terminology.

Among them, there is shown in formula (24):

$$\delta_t = \gamma \cdot 1_{\{y_t \in \mathcal{D}\}} \quad (24)$$

By statistically analyzing historical speech data, calculate the frequency and contextual importance of different vocabulary in actual use. Words that appear frequently in emergency situations are given higher priority scores. $\gamma > 0$ is the domain term enhancement coefficient; $1_{\{\cdot\}}$ is the indicator function. The composition and classification of the professional term dictionary are shown in Table 3:

Table 3. Composition and classification of professional term dictionary.

Type	Term	Part of voice	Applied context	Priority (enhancement coefficient γ)	Application scenarios
Voltage level	110kV, 220kV	Nouns/Units	Commonly seen in device recognition and operation instructions	1.2	Trigger warning mechanism
Device	Main transformer, busbar, knife switch, lightning arrester	Noun	High frequency words for operation dispatching instructions	1.4	emergency measure
Operation Action	Closing, tripping, closing, isolating	Verb	Action recognition, key instruction judgment	1.6	Condition monitoring
Status judgment	Normal, Abnormal, Fever, Alarm	Adjectives/ Status words	Key recognition in fault context	1.3	Equipment identification and operation
Direction and position	Left side, Secondary circuit, Main wiring	Noun	Electrical diagram understanding and positioning scenarios	1.1	Equipment control instructions

To improve the robustness of the model in complex and changeable voice environments, the spectrogram enhancement is combined with the background noise synthesis mechanism to achieve modeling of voice changes in dispatching scenarios. First, the signal spectrum of Wav2Vec2.0 is processed by time domain masking and frequency masking. The input spectrum tensor is represented by $S \in \mathbb{R}^{\mathrm{T}\times\mathrm{F}}$, where T represents the number of frames at the moment and F represents the dimension of the frequency. $[t, t+\omega]$ and $[f, f+h]$ are used to define the time domain masking area and the frequency domain masking area. The result after conversion is shown in formula (25):

$$S'_{i,j} = \begin{cases} 0, \text{ if } i \in [t, t+\omega] \text{ or } j \in [f, f+h] \\ S_{i,j}, \text{ otherwise} \end{cases} \quad (25)$$

In formula (25), $S'_{i,j}$ is the spectral value after occlusion. On this basis, the background noise is synthesized to construct a set of background noise sets $\mathcal{N} = \{n_1, n_2, \cdots, n_m\}$, from which a set of noise $n \sim \mathcal{N}$ is randomly selected, and the SNR (Signal-to-Noise Ratio) is used to control the mixing ratio to obtain enhanced samples, as shown in formula (26):

$$x' = x + \alpha \cdot n, \text{ with } \alpha = \sqrt{\frac{\mathbb{E}[x^2]}{\mathbb{E}[n^2] \cdot 10^{SNR/10}}} \quad (26)$$

In formula (26), x' is the speech sample after adding noise, and α is the scaling factor that controls the mixing ratio.

Based on input enhancement and domain adaptation, a joint optimization learning framework is established. The training objective is not just to rely on the traditional cross entropy loss, but on this basis, CTC loss and semantic prediction loss of the language model are applied to jointly form the end-to-end optimization objective, as shown in formula (27):

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{CTC}} + \lambda_2 \mathcal{L}_{\text{attn}} + \lambda_3 \mathcal{L}_{\text{domain}} \quad (27)$$

In formula (27), $\lambda_1, \lambda_2, \lambda_3$ are loss weights, which are selected by the hyperparameter optimization method. Among them, $\mathcal{L}_{\text{CTC}}$ is the optimization target of the frame sequence during the alignment process; $\mathcal{L}_{\text{attn}}$ is the cross entropy loss of the autoregressive decoder; $\mathcal{L}_{\text{domain}}$ is to determine the term recognition loss, which is defined as formula (28):

$$\mathcal{L}_{\text{domain}} = -\sum_{t=1}^{T} 1_{\{y_t \in \mathcal{D}\}} \cdot \log P'\left(y_t \middle| y_{<t}, X\right) \quad (28)$$

In formula (28), $1_{\{y_t \in \mathcal{D}\}}$ is the indicator function, which is 1 when the predicted word y_t belongs to the terminology dictionary $\mathcal{D}$, and 0 otherwise. Using the joint optimization method, the robust expression of multi-noise environment, the priority recognition of domain vocabulary, and the semantic modeling of context are learned simultaneously during the training process.

4. Power Grid Dispatching ASR Evaluation

A. Experimental Setting

The experimental data in this paper comes from a total of 37,144 dispatching voice calls and terminal samples collected by a municipal power grid dispatching center from 2021 to 2023. The collected original call data is about 260 hours, and the dispatching instructions cover six operating scenarios: accident handling, device switching, maintenance operations, switching operations, power generation planning, and issuance and information feedback. In terms of data preprocessing, the collected voice signals are uniformly converted into single-channel WAV with a sampling rate of 16 kHz; the signal is

downsampled using a FIR anti-aliasing filter; the silent segments are removed by joint detection of short-time energy and zero-crossing rate. On this basis, a two-stage denoising is achieved through spectral subtraction and Wiener filter, and the dynamic range of the audio is normalized. Its parameter settings are shown in Table 4:

Table 4. Preprocessing parameter settings.

Classification	Parameter	Specification
Silent detection threshold	Short-term average energy threshold	0.01
	Zero-crossing rate threshold	0.1
Noise reduction	SNR improvement target	15dB
Effective fragment screening	Minimum length	0.5 s
	Maximum mute ratio	Not exceeding 30%
RMS	Normalize target value	-26 dBFS

By preprocessing the data, 186,437 voice clips are obtained, and combined with manual annotation technology, a text alignment corpus containing 9532 professional terminology entities is established. In order to verify the generalization ability of the model and supplement the limitations of a single dataset, this paper further uses a synthetic dataset for experiments. Write scripts covering different operational scenarios (such as accident handling, equipment switching, etc.) using knowledge and practical operation processes in the field of power grid dispatching. Generate simulated dialogue audio using text to speech (TTS) technology to ensure the inclusion of specialized terminology specific to power grid dispatch. In order to enhance data diversity, pulse noise is introduced as an extreme noise condition, and four background noise environments are simulated: transformer buzzing, wind and rain noise, walkie talkie noise, and impact noise. Create different acoustic scenes through the pyroacoustics library and embed the generated speech into them to form the final synthesized audio, applying the same preprocessing steps as the original data to these synthesized audio.

To verify the effectiveness of the method proposed in this article, evaluations were conducted from several dimensions, including overall speech recognition accuracy, instruction completeness, terminology adaptation ability, noise environment adaptation ability, model efficiency, and ablation experiments. The specific parameter settings are shown in Table 5:

Table 5. Algorithm parameter settings.

Classification	Parameter	Specification
Wav2Vec 2.0	Feature encoder	One-dimensional convolution
	Masking ratio	0.065
	Masking width	10 frames
	Gumbel softmax temperature	2.0
	Negative sample size	100
	Learning rate (fine tuned)	1.00E-04
	Batch size	32
	Fine tune the number of rounds	50
Conformer	Conformer block layers	16
	Hidden layer dimension	512
	Multi head attention head count	8
	Feedforward layer dimension	2048
	Convolutional kernel size	31
	Dropout ratio	0.1
	CTC weight coefficient	0.3
	Language model predicts loss weights	0.2
	Joint optimization of total loss weight ratio	Attention: CTC: Language model = 0.5:0.3:0.2

The Conformer architecture design adopts 12 Conformer blocks and a convolution kernel size of 31. During the training process, a masking rate of 0.065 was used. This choice is based on the results of experimental tuning, and the parameter settings in this section enable the model to achieve the best balance between performance and efficiency. In the sensitivity analysis of weight coefficients, a joint loss function was used to observe the changes in model performance by fixing two weights and adjusting the third weight. Taking into account the experimental results, we ultimately chose Attention: CTC: Language model = 0.5:0.3:0.2 as the default weight configurations. This configuration achieves a good balance between recognition accuracy, terminology recognition ability, and convergence speed, which can meet the requirements of the power grid dispatch voice interaction system.

To present the experimental results, the following models are selected for comparison:

DNN-HMM (Deep Neural Network - Hidden Markov Model): a classic hybrid acoustic model framework that uses a deep feedforward neural network to establish the state of the HMM.

DeepSpeech2: an end-to-end ASR model based on CNN and bidirectional recurrent neural network, which applies CTC loss for optimization and is robust and practical.

Transformer ASR: using its own attention mechanism, a complete sequence model is built to achieve end-to-end ASR, which can obtain remote context information and has strong semantic modeling capabilities.

B. Experimental Results

1) Overall Voice Recognition Accuracy

This paper uses WER and CER as the main evaluation indicators, selects 20% of the original corpus from 6 dispatching scenarios as test samples, and compares the overall recognition accuracy of each model. The results are shown in Figure 3:

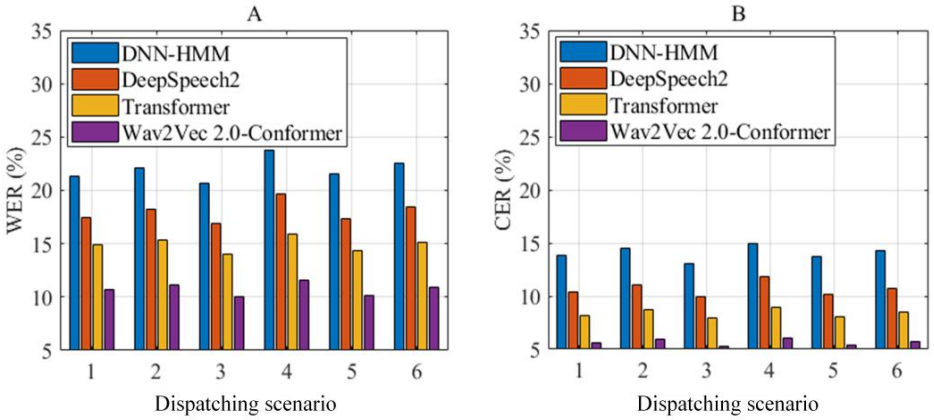

Figure 3. Recognition accuracy comparison. Figure 3A shows the WER result; Figure 3B shows the CER result.

From Figure 3, the model based on Wav2Vec 2.0 and Conformer structure proposed in this paper shows significant advantages in WER and CER results in 6 types of power grid dispatching scenarios. In Figure 3A, the average WER of this model in the dispatching scenario reaches 10.8%, while the average WER of DNN-HMM, DeepSpeech2, and Transformer ASR reaches 22.0%, 18.0%, and 14.9%, respectively. Compared with the comparison model, the average WER of this model is reduced by 11.2%, 7.2%, and 4.1%, respectively. In Figure 3B, the average CER of this model is 5.7%, while the average CER of DNN-HMM, DeepSpeech2, and Transformer ASR is 14.1%, 10.7%, and 8.4%, respectively. Compared with the comparison model, the average CER of this model is reduced by 8.4%, 5.0%, and 2.7%, respectively.

This result shows that the proposed model has higher accuracy in overall semantic recognition. In the pre-training stage, Wav2Vec2.0 is used to fully explore the contextual information in the original audio and improve its representation ability, so that it can still have high accuracy in complex dispatching environments. The Conformer structure improves the model's perception of local voice changes. In comparison, the DNN-HMM model is difficult to adapt to unstructured voice instructions due to its weak feature representation ability; DeepSpeech2 and Transformer have strong modeling depth, but still face generalization bottlenecks when processing highly domain-specific terms.

To more precisely reflect the accuracy of the model for high-risk voice instructions in power grid dispatching, the Weighted Word Error Rate (W-WER) indicator is applied. For voice instructions with greater operational significance and stronger safety, such as "tripping", "load", "main transformer", "decoupling", and "busbar", high penalty weights are given to reflect their "semantic priority" in actual dispatching. In the experimental set, 2,368 corpora containing at least one type of key voice instructions are selected. According to the importance of voice instructions in the dispatching process, High-Risk instruction (weight set to 3.0) and Medium-Risk instruction (weight set to 2.0) are divided into two categories. On this basis, the W-WER of each model is calculated and compared with the traditional WER, as shown in Table 6.

Table 6. Comparison of different models in standard WER and weighted WER.

Model	WER (%)	W-WER (%)	High-Risk instruction recognition rate (%)	Medium-Risk instruction recognition rate (%)
DNN-HMM	22.0	28.4	71.2	79.5
DeepSpeech2	18.0	23.6	76.8	82.3
Transformer ASR	14.9	19.1	83.4	86.7
Wav2Vec2.0-Conformer	10.8	13.8	91.5	92.6

From Table 6, the proposed model is significantly better than other models in both standard WER and weighted WER, which are 10.8% and 13.8%, respectively, and the recognition rate of High-Risk instruction is 91.5%. DeepSpeech2 and Transformer ASR are 23.6% and 19.1% in W-WER respectively, which shows that the model still has the problem of insufficient semantic concentration when processing High-Risk instruction. The weighted word error rate of the DNN-HMM model is 28.4%, indicating that the model has weak recognition ability for key instructions and has a large risk of misrecognition and missed detection. Overall, the proposed model can more accurately extract the features of voice instructions and can better improve the system's overall recognition accuracy.

2) Instruction Completeness

To evaluate the model's ability to understand and restore complete dispatching statements, the instruction completeness recognition rate is used as a measurement indicator. Combined with the average real-time factor (RTF) for comprehensive evaluation, the model's accurate recognition ratio of complete dispatching instructions containing key terms is analyzed. A total of 1,573 complete dispatch instruction samples containing at least two key terms are selected from the dataset. Each instruction contains a subject (device name), a predicate (operation action), and additional conditions (voltage level, time limit). The comparison results are shown in Table 7.

Table 7. Instruction completeness comparison.

Model	Complete recognition rate (%)	RTF	Complete recognition of high-risk instruction numbers
DNN-HMM	63.8	0.21	521
DeepSpeech2	71.5	0.24	658
Transformer ASR	77.2	0.28	743
Wav2Vec2.0-Conformer	84.6	0.17	817

As can be seen from Table 7, the complete recognition rate of the DNN-HMM model is the lowest, only 63.8%, which indicates that the model cannot extract the complex relationship between multiple instructions well and is prone to misrecognition. Although DeepSpeech2 and Transformer ASR have improved on this indicator compared with DNN-HMM, with complete recognition rates of 71.5% and 77.2%, respectively, they are still inferior to the model in this paper. The complete recognition rate of the model in this paper reaches 84.6%, and the RTF is 0.17, which is better than the other three models. It can more completely recognize the logical relationship between consecutive words based on real-time interaction needs.

3) Terminology Adaptation Ability

The terminology adaptation ability test analyzes the adaptation ability of different models to professional terms through three dimensions: term recognition precision, recall rate, and F1 value. The evaluation corpus is based on the existing vocabulary alignment corpus. By aligning the recognition results of each model with the existing labeled vocabulary, the number of correctly recognized terms, missed detections, and misrecognitions are counted, and the evaluation indicators are calculated based on these data. The final results are shown in Table 8:

Table 8. Terminology adaptation evaluation.

Model	Precision (%)	Recall (%)	F1 Score (%)
DNN-HMM	81.2	75.6	78.3
DeepSpeech2	86.7	82.1	84.3
Transformer ASR	88.4	84.5	86.4
Wav2Vec2.0-Conformer	92.3	89.1	90.7

From the evaluation results in Table 8, compared with the control model, the proposed method has greatly improved in precision, recall, and F1 value, with specific results of 92.3%, 89.1%, and 90.7%, respectively, and improved by 3.9%, 4.6%, and 4.3%, respectively compared with the second-best Transformer ASR. The DNN-HMM model scores the lowest in the three indicators, with an F1 value of only 78.3%, indicating that the model has certain limitations in its ability to extract professional terms in complex scenarios.

The DNN-HMM model is difficult to achieve efficient modeling of complex semantic relationships and has poor adaptability to term deformation and non-standard voice, resulting in low precision and recall. Although DeepSpeech2 has shown some advantages in the end-to-end framework, the generalization performance of vocabulary expression is still insufficient in the absence of training samples. Transformer ASR has good modeling capabilities, but due to the lack of voice feature perception modules, its advantages in capturing local phoneme information are not obvious enough. This paper combines Wav2Vec2.0 with Conformer to extract rich expression information from massive unlabeled voice data in a self-supervised manner and uses Conformer to deeply model high-dimensional scenario expressions, so

that it has stronger modeling adaptability in power grid dispatching corpus with high concentration of professional terms.

To further evaluate the model's real adaptability in the case of fuzzy pronunciation of key terms in each scenario or the presence of oral variants, the professional term Top-K recognition rate is further compared. The coverage is calculated by extracting the Top-3 candidate outputs corresponding to each voice and comparing them with the term label words. The final result is shown in Figure 4:

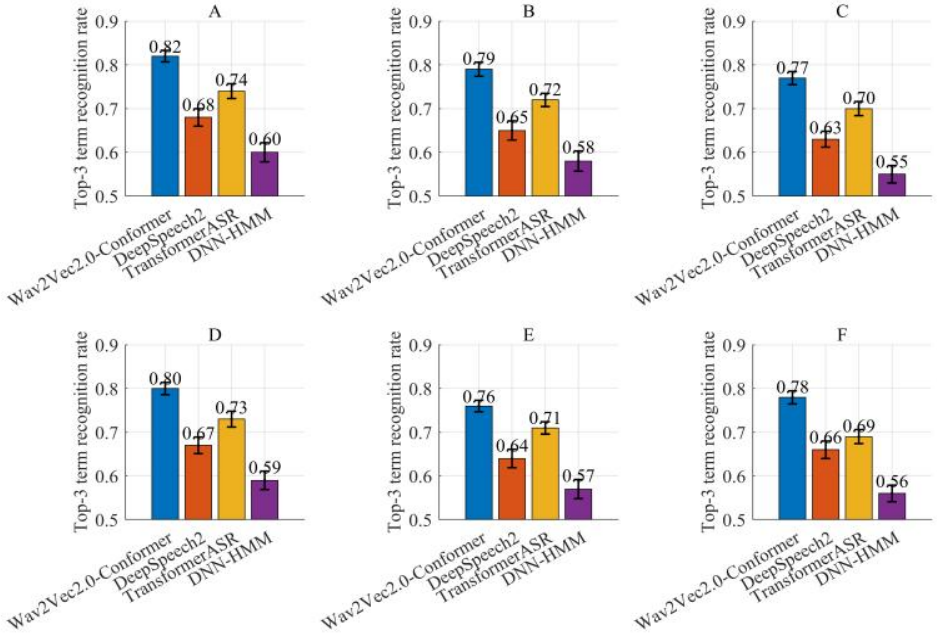

Figure 4.Top-3 recognition rate comparison. Figure 4A is the accident handling recognition rate; Figure 4B is the device switching recognition rate; Figure 4C is the maintenance operation recognition rate; Figure 4D is the switching operation recognition rate; Figure 4E is the power generation plan recognition rate; Figure 4F is the issuance and information feedback recognition rate.

As can be seen from Figure 4, the model in this paper has the highest recognition rate in each dispatching scenario, and the error is small, which means that it has good adaptability. From Figure 4A, Figure 4B, Figure 4C, Figure 4D, Figure 4E, and Figure 4F, the Top-3 recognition rate of the term in each dispatching scenario of the model in this paper is above 0.75. This is because Wav2Vec2.0 has the ability of self-supervised learning, which can better mine the deep features of voice and improve the contextual understanding and recognition ability of specific domain vocabulary through the efficient time series modeling mechanism of Conformer. It can still extract the main category features well in the complex scenario with large changes in power grid dispatching.

Although Transformer ASR has strong global context modeling ability, its adaptation under small sample conditions is still limited, making the Top-3 recognition rate lower than that of the model in this paper. DeepSpeech2 and DNN-HMM have relatively simple structures. The former relies on CTC loss, and the network has limited ability to capture long-distance dependencies; the latter designs the acoustic model and language model separately, resulting in poor performance in complex vocabulary recognition. Overall, the model in this paper integrates pre-training and high-level time series structure, which can effectively improve the adaptability of professional terms in power grid dispatching.

On this basis, more competitive ASR benchmark models, Whisper and HuBERT, were introduced in the experiment. Whisper is a large-scale pre trained ASR model proposed by OpenAI, while HuBERT is a self supervised learning model proposed by Meta. This article compares the recognition accuracy (experimental results) and prediction accuracy (model predicted values) of different models on a dataset. As shown in Table 9:

Table 9. Comparison of advanced model experimental results and predicted values.

Model	Experimental accuracy (%)	Prediction accuracy (%)
Whisper	88.7	88.1
HuBERT	89.1	88.4
Wav2Vec2.0-Conformer	90.6	90.2

From Table 9, it can be seen that the prediction accuracy based on Wav2Vec 2.0 and Conformer model is very close to the experimental accuracy, with a deviation of less than 0.5%. This indicates that the model has high

prediction accuracy and stability. Especially in the dataset, its prediction accuracy reached 90.2%, and the experimental accuracy was 90.6%, further verifying the effectiveness of the model. Although Whisper and HuBERT performed relatively well on the dataset, the recognition accuracy of our model was slightly higher than these two models on all datasets, with experimental accuracies 1.9% and 1.5% higher than Whisper and HuBERT, respectively. This indicates that the model presented in this article has higher accuracy in handling complex speech interaction tasks.

4) Adaptability to Noise Environments

To verify the robustness of the model under background noise, this paper applies impulse noise as an extreme noise condition and uses the SNR robustness curve to compare the performance of various models under multiple background noise environments, setting the signal-to-noise ratio to -5-50 dB. Under different signal-to-noise ratios, the WER of each model is calculated, and the results are shown in Figure 5.

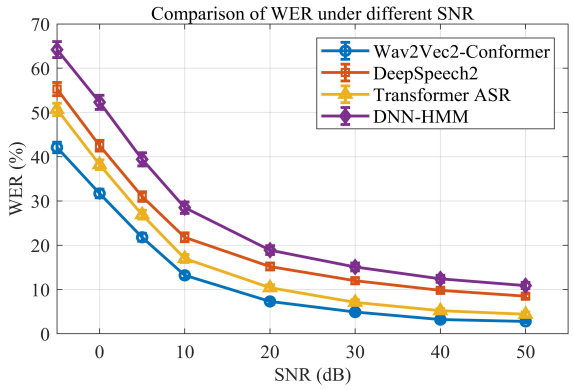

Figure 5. Comparison of adaptation to noise environment.

In Figure 5, Wav2Vec2-Conformer has the best robustness to impulse noise. When the signal-to-noise ratio reaches -5dB, the WER can still be controlled at 42.1%, which is significantly better than other models, and as the SNR increases, the WER continues to decrease, reaching only 2.8% at 50 dB. This shows that the proposed model can more efficiently extract noise-resistant robust features by combining context-aware voice features with the Conformer structure with good local and global modeling capabilities. The performance of DNN-HMM and DeepSpeech2 at a signal-to-noise ratio of -5 dB is not ideal, and their WERs are significantly higher, with limited feature extraction and insensitivity to background noise. The performance of Transformer ASR in the low signal-to-noise ratio area is still somewhat different from that of the proposed model. This is because Transformer ASR has strong global modeling capabilities and is easily affected by external noise.

To be closer to the actual operation of the power system, the recognition robustness of the model under typical power grid noise conditions is evaluated. This paper selects four types of noise: transformer hum, wind and rain noise, intercom noise, and communication channel interference. Under the same signal-to-noise ratio (SNR=10dB), the WER of each model is statistically analyzed, and the results are shown in Figure 6:

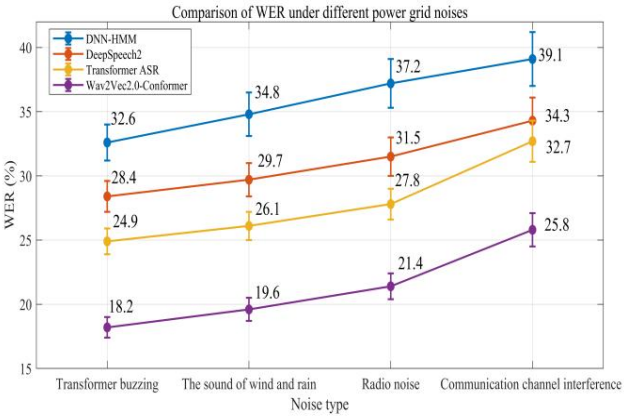

Figure 6. Comparison of WER under different noise conditions.

Under the four typical power grid noise conditions in Figure 6, the proposed model is significantly better than the DNN-HMM, DeepSpeech2, and Transformer ASR comparison models in terms of WER. The experimental results show that the WER of the proposed model under various types of noise are 18.2%, 19.6%, 21.4%, and 25.8%, respectively. Based on the Wav2Vec 2.0 and Conformer fusion model, the spectrogram enhancement is combined with the background noise synthesis mechanism, which shows stronger anti-interference ability in relatively stable noise environments such as transformer hum and still maintains good recognition stability under more challenging communication channel interference, further improving the model's anti-interference ability to noise, thereby effectively improving recognition robustness.

5) Model Efficiency

In order to comprehensively evaluate the performance of the proposed model and the comparative model in terms of model efficiency, a computational resource consumption index was used to record the GPU usage and average processing time of each instruction for each model, as shown in Table 10:

Table 10. Comparison of model efficiency.

Model	GPU usage rate (%)	Average instruction processing time (ms)
DNN-HMM	65.2	120.8
DeepSpeech2	68.9	105.2
Transformer ASR	72.1	93.7
Wav2Vec2.0-Conformer	61.5	81.1

From Table 10, it can be seen that although DNN-HMM has relatively low computational resource consumption, its processing speed is slow. DeepSpeech2 and Transformer ASR show high GPU utilization and fast processing speed respectively, but due to their complex network structure and a large number of parameters, they may face higher hardware requirements in actual deployment. In contrast, the model presented in this article demonstrates more reasonable GPU utilization while maintaining efficient processing speed. Especially in terms of average processing time for each instruction, it is only 81.1 milliseconds, significantly better than the other three models.

6) Ablation Experiment

To verify the specific contribution of each key module in the model proposed in this paper to the overall performance, the Wav2Vec2.0 pre-trained encoder, Conformer structure, and term-guided attention mechanism are eliminated, and the WER and CER results of each ablation model are compared to verify the specific contribution of each key module in the model to the overall performance. The results are shown in Figure 7.

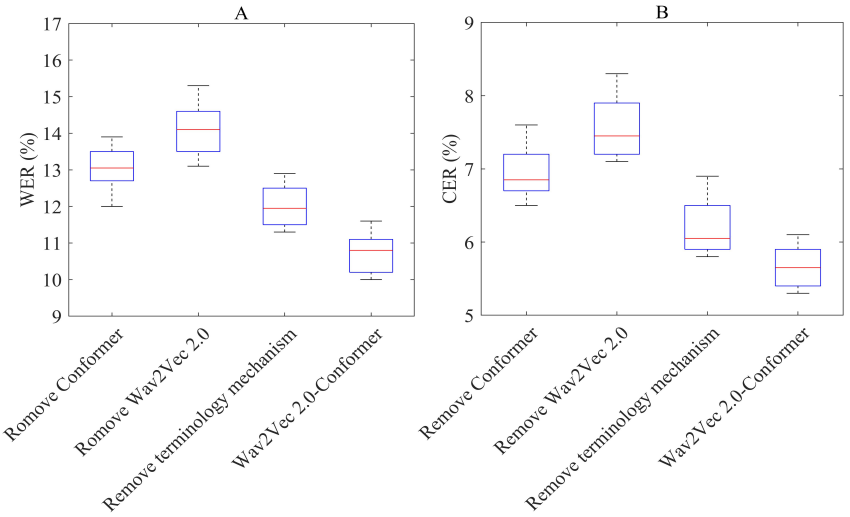

Figure 7. Comparison of ablation experiment results. Figure 7A is the WER ablation experiment result; Figure 7B is the CER ablation experiment result.

In Figure 7, there is a large difference between the WER and CER results after eliminating each component. From Figure 7A and Figure 7B, after removing the Wav2Vec2.0 pre-trained encoder, the model performance is greatly reduced; the WER is increased to 13.1%-15.3%; the CER is increased to 7.1%-8.3%. After removing the Conformer structure, the ratio of WER to CER also increases significantly, 12.0%-13.9% and 6.5%-7.6%, respectively, indicating that Wav2Vec2.0 and Conformer have a significant impact on temporal correlation and local overall characteristics, while the absence of them reduces the model's performance. After removing the term-guided attention mechanism, WER and CER increase to 11.3%-12.9% and 5.8%-6.9%, respectively, which is slightly lower than the complete model, indicating that this mechanism helps to improve the term recognition ability and context focus ability. In general, the three modules play an important role in improving the accuracy of ASR, which means that the design of each component in this paper is reasonable and has a synergistic effect.

5. Conclusions

The existing power grid dispatching voice interaction system has the problem of difficulty in recognizing professional terms and poor scenario adaptability. To improve the accuracy and robustness of dispatch instruction ASR, this paper combines Wav2Vec 2.0 and Conformer to build an end-to-end ASR model. Through Wav2Vec 2.0 self-supervised pre-training and Conformer structure modeling, the terminology-guided attention mechanism is integrated to improve the recognition ability of professional terminology. The experimental results show that the overall recognition accuracy of this model is significantly better than DNN-HMM, DeepSpeech2, and Transformer ASR, with an average WER and CER of 10.8% and 5.7%, respectively; the F1 value of term recognition reaches 90.7%, showing strong robustness, and the WER decrease trend is better. The ablation experiment verifies the key role of each key component in improving the model's performance, which provides a certain reference for solving the problems of professional terminology adaptation and noise robustness in power grid dispatch ASR. However, this paper still has limitations. This paper does not deeply explore the adaptability of the model to extreme noise and rare accents. In practical operation, it shows a certain demand for computing resources. According to the experimental test results, the average GPU utilization of the system is 61.5%, and the average delay of a single speech processing is 81.1ms. Although it consumes less computing resources compared to other models, it still has a certain delay on low performance hardware. In addition, this article did not conduct in-depth research on user interaction experience, nor did it effectively explore the design of voice interaction interfaces and the optimization of command input processes. Future research can focus on multimodal integration,

lightweight model design, and dynamic adaptive mechanisms. Speech enhancement techniques based on deep learning can further improve noise reduction capabilities. The system can be deployed in the server cluster of the power grid dispatching center. Multiple terminal devices can be connected through the LAN and deployed locally using edge computing devices to further improve the practicability and intelligence of the power grid dispatching ASR system.

Acknowledgment

None

Consent to Publish

The manuscript has neither been previously published nor is under consideration by any other journal. The authors have all approved the content of the paper.

Funding

None

Author Contribution

[Min Gao]: Developed and planned the study, performed experiments, and interpreted results. Edited and refined the manuscript with a focus on critical intellectual contributions.

[Chenguang Zhu, Lei Chen]: Participated in collecting, assessing, and interpreting the date. Made significant contributions to date interpretation and manuscript preparation.

[Weizhe Sun, Wengang Wang]: Provided substantial intellectual input during the drafting and revision of the manuscript.

Conflicts of Interest

The authors declare that they have no financial conflicts of interest.

References

[1] Y.L. Fan, H. Wang, Y.D. Bai, X. Li. Research on power grid dispatching and control business based on artificial intelligence technology. Power Systems and Big Data, 2020, 23(5), 9-15.

[2] S.X. Fan, J.B. Guo, S.C. Ma, L.X. Li, G.Z. Wang, H.T. Xu, et al. Framework and Key Technologies of Human-machine Hybrid-augmented Intelligence System for Large-scale Power Grid Dispatching and Control. CSEE Journal of Power and Energy Systems, 2024, 10(1), 1-12. DOI: 10.17775/CSEEJPES.2023.00940

[3] S.B. Zeng, D.K. Hong, F.F. Hu, L. Liu, F. Xie. Application of intelligent speech analysis based on BiLSTM and CNN dual attention model in power dispatching. Nanotechnology for Environmental Engineering, 2021, 6(3), 1-8. DOI: 10.1007/s41204-021-00148-7

[4] J.J. Liu, A. Wumaier, D.P. Wei, S. Guo. Automatic speech disfluency detection using wav2vec2. 0 for different languages with variable lengths. Applied Sciences, 2023, 13(13), 7579. DOI: 10.3390/app13137579

[5] J. Seo, B. Lee. Multi-task conformer with multi-feature combination for speech emotion recognition. Symmetry, 2022, 14(7), 1428-1438. DOI: 10.3390/sym14071428

[6] P. Zhao, F.G. Liu, X.Q. Zhuang. Speech sentiment analysis using hierarchical conformer networks. Applied Sciences, 2022, 12(16), 8076-8091. DOI: 10.3390/app12168076

[7] Q. Zhao, T.R. Li, R. Luo, R. Li, T.Y. Han, D.S. Han. Power dispatch speech recognition based on dual dictionary class label language model. Electronic Measurement Technology, 2021, 44(13), 121-126. DOI: 10.19651/j.cnki.emt.2106694

[8] L. Xiao, Q.H. Xiao, L.L. Wei, Y.Y. Zhou, S. Wang. Research on acoustic model of power grid dispatch speech recognition based on big data and deep learning. Power Systems and Big Data, 2022, 25(9), 30-36.

[9] Z.H. Wang, F. Gao. Research on voice interaction model of intelligent power dispatching based on DCGAN. Nanotechnology for Environmental Engineering, 2021, 6(3), 53-60. DOI: 10.1007/s41204-021-00150-z

[10] L. Chen, W.Y. Zheng, H.H. Yu, J. Fu, H.W. Liu, J.Q. Xia. Research on BERT-based language model for power grid dispatch speech recognition. Power System Technology, 2020, 45(8), 2955-2961. DOI: 10.13335/j.1000-3673.pst.2020.0796

[11] F. Hao, X.H. Wang, C.J. Pang. Word2vec-based method for generating word vectors for power grid dispatching and its application in speech recognition. Inner Mongolia Electric Power, 2020, 38(5), 72-76. DOI: 10.3969/j.issn.1008-6218.2020.00.076

[12] L.L. Sun, Q. Zhai, Y.T. Zhang, H.T. Zhai, Q.R. Zhang. Design of power grid dispatching authentication system based on voiceprint recognition. Shandong Electric Power, 2020, 50(10), 58-65. DOI: 10.20097/j.cnki.issn1007-9904.2023.10.008

[13] J.H. Geng, D.Y. Jia, Z.H. He, N.K. Wu, Z.Q. Li. Enhanced Conformer-Based Speech Recognition via Model Fusion and Adaptive Decoding with Dynamic Rescoring. Applied Sciences, 2024, 14(24), 11583-11608. DOI: 10.3390/app142411583

[14] Z.P. Fan, X.J. Zhang, M. Huang, Z.H. Bu. Sampleformer: An efficient conformer-based Neural Network for Automatic Speech Recognition. Intelligent Data Analysis, 2024, 28(6), 1647-1659. DOI: 10.3233/IDA-230612

[15] X. Zhang, X.C. Zhang, W.S. Chen, C.L. Li, C.Y. Yu. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. Scientific Reports, 2024, 14(1), 9543-9556. DOI: 10.1038/s41598-024-60278-1

[16] A. Baevski, H. Zhou, A. Mohamed, M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, 12449-12460. DOI: 10.48550/arXiv.2006.11477

[17] J. Cai, Y.L. Song, J.H. Wu, X. Chen. Voice disorder classification using Wav2vec 2.0 feature extraction. Journal of Voice, 2024. DOI: 10.1016/j.jvoice.2024.09.002

[18] S. Axelrod, R. Gomez-Bombarelli. Molecular machine learning with conformer ensembles. Machine Learning: Science and Technology, 2023, 4(3), 035025. DOI: 10.1088/2632-2153/acefa7

[19] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J.R. Hershey, L. Jones, et al. DF-Conformer: Integrated architecture of Conv-TasNet and Conformer using linear complexity self-attention for speech enhancement. 2021

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2021. DOI: 10.48550/arXiv.2106.15813

[20] J. Zhao, W.Q. Zhang. Improving automatic speech recognition performance for low-resource languages with self-supervised models. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6), 1227-1241. DOI: 10.1109/JSTSP.2022.3184480

[21] B. Deng, C.Y. Peng, Z.B. Zhang. Research on speech recognition for smart grid dispatch based on Conformer model. Manufacturing Automation, 2024, 46(6), 126-131. DOI: 10.3969/j.issn.1009-0134.2024.06.020

[22] J.K. Sang, N. Yuruvas. Compression optimization strategy for end-to-end speech recognition model based on Conformer. Journal of Signal Processing, 2022, 38(12), 2639-2649. DOI: 10.16798/j.issn.1003-0530.2022.12.018

[23] M. Vaithianathan. Digital Signal Processing for Noise Suppression in Voice Signals. International Journal of Advanced Research and Interdisciplinary Scientific Endeavours, 2024, 1(4), 198-208. DOI: 10.61359/11.2206-2417

[24] C.S. Zheng, X.H. Hu, Y. Zhou, X.D. Li. Spectral subtraction based on noise spectral structure characteristics. Acta Acustica, 2022, 35(2), 215-222. DOI: 10.15949/j.cnki.0371-0025.2010.02.020

[25] R.H. Cao, X.L. Wu, C. Feng, F. Zheng, M.X. Xu. Emotion recognition of conversational speech based on Wav2vec2.0 and contextual emotion information compensation. Journal of Signal Processing, 2023, 39(4), 698-707. DOI: 10.16798/j.issn.1003-0530.2023.04.011

[26] Q.S. Zhu, J. Zhang, Y. Gu, Y.C. Hu, L.R. Dai. Multichannel av-wav2vec2: A framework for learning multichannel multi-modal speech representation. In Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17), 19768-19776. DOI: 10.1609/aaai.v38i17.29951

[27] M.F. Ghobrial, A.M. Gody, S.T. Muhammad. Comparative Study on End-to-End Speech Recognition Using Pre-trained Models. Fayoum University Journal of Engineering, 2025, 8(1), 131-142. DOI: 10.21608/fuje.2024.312102.1089

[28] C.F. Gao, G.F. Cheng, P.Y. Zhang. A consistent self-supervised learning approach for robust automatic speech recognition. Acta Acustica, 2023, 48(3), 578-587. DOI: 10.15949/j.cnki.0371-0025.2023.03.008

[29] A. Mohamed, H. Lee, L. Borgholt, J.D. Havtorn, J. Edin, C. Igel. Self-supervised speech representation learning: A review. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6), 1179-1210. DOI: 10.1109/JSTSP.2022.3207050

[30] X.H. Yue, J.R. Lin, F.R. Gutierrez, H.Z. Li. Self-supervised learning with segmental masking for speech representation. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6), 1367-1379. DOI: 10.1109/JSTSP.2022.3191845

[31] A. Baevskii, H. Zhou, A. Mohamed, M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, 12449-12460. DOI: 10.48550/arXiv.2006.11477

[32] Y.T. Li, D. Qu, X.K. Yang, H. Zhang, X.L. Shen. An improved linear attention mechanism for speech recognition. Journal of Signal Processing, 2023, 39(3), 516-525. DOI: 10.16798/j.issn.1003-0530.2023.03.014

[33] F.F. Wang, K.R. Ben, X. Zhang. Research on robustness enhancement technology for speech recognition based on domain knowledge. Computer Engineering & Science, 2023, 45(12), 2155-2164. DOI: 10.3969/j.issn.1007-130X.2023.12.007

[34] X.H. Huang, L.S. Qiao, W.T. Yu, J. Li, Y.Z. Ma. End-to-end sequence labeling via convolutional recurrent neural network with a connectionist temporal classification layer. International Journal of Computational Intelligence Systems, 2020, 13(1), 341-351. DOI: 10.2991/ijcis.d.200316.001